Design and Development of Machine Learning Model for Multi-Factor Dependent Prediction and Analysis

SYNOPSIS

Submitted to

SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY, NANDED

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Subject of

Computer Science

Under the Faculty of Science and Technology

Submitted by

Mr. Suryakant Vishwasrao Patil

Under the Supervision of

Dr. Renuka R. Londhe

Assistant Professor and HoD,

Rajarshi Shahu Mahavidyalaya (Autonomous), Later

Research Centre

College of Computer Science and IT, Latur 413512

March 2022

Contents

Abstract:.....

1. Introduction: 1
2. Literature review: 2
3. Research Objectives: 6
4. Methodology to be used: 6
5. Importance of Study/society Applications: 10
6. Proposed Work Plan: 10
7. References 11

Abstract:

Artificial Intelligence (AI) and Machine Learning have gained widespread acceptance in recent years as a result of their practical applications in areas such as automated vehicles, intelligent robots, and image processing. In the fields of prediction, analysis, and forecasting, there are numerous AI-based applications. Using Machine Learning approaches, a lot of studies have been done on predicting and analyzing outputs on massive datasets (ML). We will be using a benchmark study to compare the performance of various machine learning algorithms on three separate datasets with varied input elements in this work. We will suggest the explainable machine learning model, which employs multiple regression and Artificial Neural Network (ANN) to predict and evaluate procedures on multi-factor inputs. The developed model will be compared to the results of existing Machine Learning Models. The purpose of multiple regression is to determine the interdependency of the factors or variables.

The input variables have a complex impact on prediction; hence it requires non-linearity in the model. Due to the introduction of non-linearity in the model, the model will be a custom design having multiple layers with non-linear activation functions in between. The proposed custom model will enhance the efficiency and effectiveness of accurate predictions. To train and test the model, we may use tools like SHAP or LIME.

1. Introduction:

Machine Learning (ML) has a wide range of industrial applications that are likely to increase in the coming areas. The global Machine Learning market is expected to reach USD 117 billion by 2027 with an impressive CAGR (Compound Annual Growth Rate) of 39%. With the use of regression methods in Machine Learning, future values can be forecasted. Regression is used to forecast a wide range of future values using the input data/factors. The label is the target variable in machine learning, and regression is used to define the relationship between the label and data. In machine learning, regression is a sort of supervised learning that aids in the mapping of a prediction relationship between labels and data. In machine learning, the most common forms of regression algorithms are linear, polynomial, logistic, stepwise, and so on.

Polynomial regression includes non-linear regression. It's a technique for simulating a non-linear relationship between dependent and independent variables. It's utilized when the data has a curved trend and linear regression would not generate as accurate findings as non-linear regression would. This is because the data in linear regression is considered to be linear.

Machine-learning approaches have been employed for prediction in sectors such as stock selection, weather forecasting, and agricultural production forecasting in recent years. There are still some unanswered problems about which method is better or most appropriate for multi-factor dependent prediction and analysis. This research intends to provide a framework for predicting the relationship between independent factors and excess returns using several ways, conduct a systematic test, and determine which method is the most effective.

The goal of this study is to identify existing prediction and analysis approaches. First, prior researchers' methodologies will be compared. A relative analysis will be conducted to determine the importance of each technique and to provide a proposal for further improvement in prediction and analysis.

Due to the introduction of non-linearity in the model, the model will be a custom design having multiple layers with non-linear activation functions in between. The proposed custom model will enhance the efficiency and effectiveness of accurate predictions and analysis.

Factor Variable Selection:

The essence of the multifactor model is to build an optimal prediction through factor selection. Therefore, factors should be selected as many as possible. The independent variables in the model, are determined by experience.

2. Literature Review:

Mark A kvamev presented a nonlinear principal component analysis for multiple-variant data Analysis. The NLPCA models successfully reduce dimensionality and produce a feature space map resembling the actual distribution of underlining system parameters [1].

Mantas Marciukitis explained the use of a nonlinear regression model for wind turbine power curves [2].

Jixiang Sun designed an investment model based on TabNet. His suggested model outperforms LSTM and GRU models by having better interoperability [3].

Jieying Gao presented a multi-factor stock selection strategy based on machine learning evidence from China [4].

Fraud Noman designed an autoregressive neural network for predicting wind speed. Used models are a) Feed Forward Neural Network b) Time Delay Neural Network c) Nonlinear autoregressive exogenous models [5].

Roohollah Shirani Faradonveh developed genetic programming and nonlinear multiple regression techniques to predict back break in blasting operation. He finds out that the genetic programming approach higher reliable as coma pairs to the nonlinear regression model [6].

Julie A. Shortridge presented how climate, agriculture, and hunger are related to each other by using nonlinear regression and data mining techniques [7].

Sambeet Mishra compared deep learning models for multi-variant prediction of time series wind power generation and temperature. The result demonstrates that attentions and DCN performed better with wavelet signals [8].

Jan Adamowski compared multiple and nonlinear regression methods for urban water demand forecasting in Mon treat Canada. The wavelet ANN models were found to be better than MLR, MNLR, and ARIMA models [9].

Jan Adamowski, Hiu Fung Chan, Shiv O. Prasher, Bogdan Ozga Zielinski, and Anna Sliusarieva, Compared multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network for methods for urban water demand forecasting in Montreal **Canada** [10].

Theresa Hoang Diem Ngo, La Puente, and CA discussed steps that should be followed in Multiple Regression Analysis [11].

Konstantinos G. Liakos reviewed research dedicated to applications of machine learning in agricultural production systems including yield prediction, disease prediction, etc. [12].

Girish L. used machine learning to predict the relation between crop yield and rainfall [13]

S. Thuankaewsing studied the harvest scheduling problem of sugarcane growers of Thailand [14].

Manoj G.S. analyzed and predicts the crop yield using machine learning techniques. The model suggests the best suitable crop [15].

Y. L. Everingham used data mining approaches to forecast regional sugar cane crop production. He proved ensemble approach using a forward stage-wise algorithm is an effective and efficient way to select and weight model for yield prediction [16].

Phusanica Charoen used a random forest model to predict yield using forward feature selection and hyper-parameter tuning [17].

Rajeshwari C carried out a comparative study of big data analytics tools like R and Tableau finds that Tableau is better than R [18].

Jasper E surveys semi-supervised learning where data is labeled or unlabeled [19].

Joachim Harvest discusses on use of the machine learning model in workflow management. The study shows how sequential workflow management is different from it [20].

Sunil Kumar uses an artificial neural network model to forecast sugar cane yield. The paper is about the development of an ANN-based model for predicting sugar cane yield in India [21].

Julie Wang defined an intelligent optimization-based multifactor deep learning stock selection model and quantitative trading strategy [22].

Devon L Robertson designed a model to predict the density of serious crime incidents using multiple input hidden Markov Maximization a posteriori model [23].

Haoxuan Li overviewed machine learning for stock selection based on multi-factor models He tested six algorithms and finds out their merits and demerits of them [24].

3. Research Objectives:

1. To apply the existing algorithms to the prepared database and evaluate the results
2. To develop a custom model and apply it to the database and compute the results.
3. To do a comparative analysis of existing algorithms and custom models.
4. To do efforts for the improvement of performance and accuracy of a custom model.

4. Methodology to be used:

The proposed machine learning techniques comprise multiple input parameters. The following fig. figure 1 illustrates the key steps for designing and developing a machine learning model for multi-factor dependent prediction and analysis.

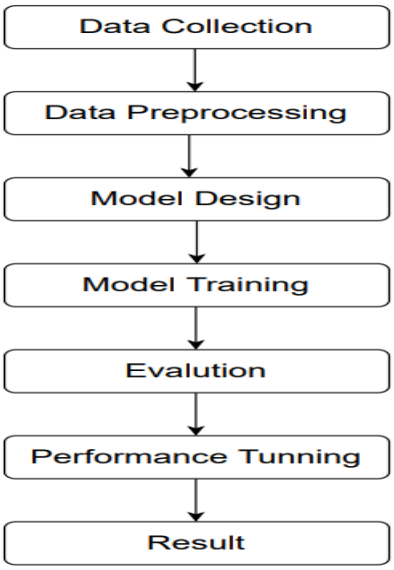


Figure1: Steps to be followed for designing and developing of machine learning model

The machine learning techniques steps are followed:

4.1 Data Collection:

The data will have to be collected from different sources in various formats for multiple input factors. Three different data sets will be collected.

4.2 Data Pre-processing

The data analysis is fully dependent on a data feed to the machine. Sometimes, the data is incomplete and inconsistent and you need to pre-process your data by normalizing, eliminating duplicates, and making error corrections.

4.3 Model Design

There are several models available for data processing but as per the optimization and multiple independent input variables involved, we will have to design a new model which will analyze the factors which are having more impact or no impact on the prediction. The following fig. figure 2 shows multiple factors involved in prediction and analysis.

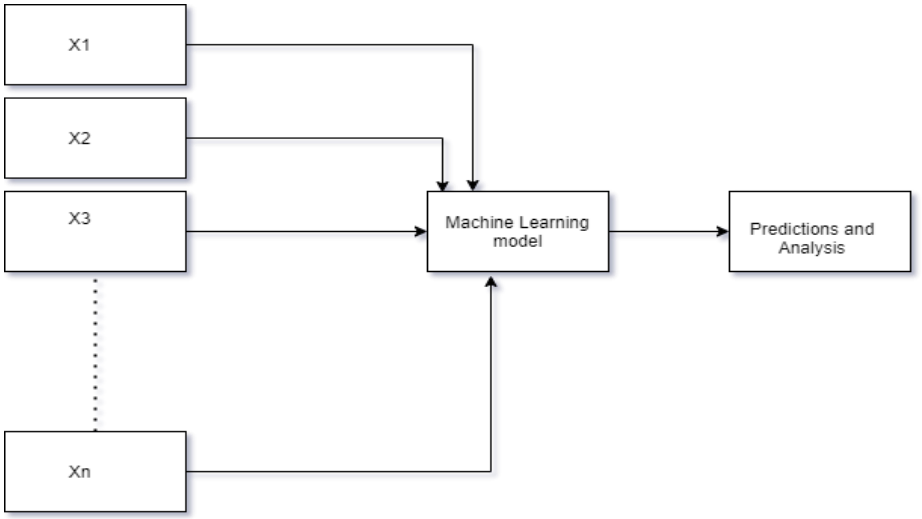


Figure 2: Multiple factors involved in prediction and analysis

Machine learning techniques such as regression analysis, and explainable machine learning will have to use for building the machine model [25-26].

Machine learning gives a better result when the number of variables increases. Regression analysis is the most powerful tool which is used for statistical analysis. Due to the number of variables, we will use multiple regression which predicates the reliability of the model and which variables have more or no impact on analysis [25].

Explainable Artificial Intelligence (XAI) is incorporated into the machine learning algorithms to trust results and output through the set of processes and methods. The machine-learning model is described by explainable AI to predicate impact and potential partialities. In

addition, it characterizes model accuracy, fairness, transparency, and outcomes, which will help in AI-powered, decision-making. Explainable AI plays a crucial role in building trust and confidence when a model is deployed in the organization as well as helps to adopt a responsible approach to develop a model [26].

The custom model is designed with a combination of regression and an Artificial Neural Network.

Methodology to be adopted:

Database:

a. Data Set 1 (Multi-factor dependent stock data).

b. Data Set 2 (Multi-factor dependent crop yield data).

c. Data Set 3 (Multi-factor dependent disease data).

Techniques adopted:

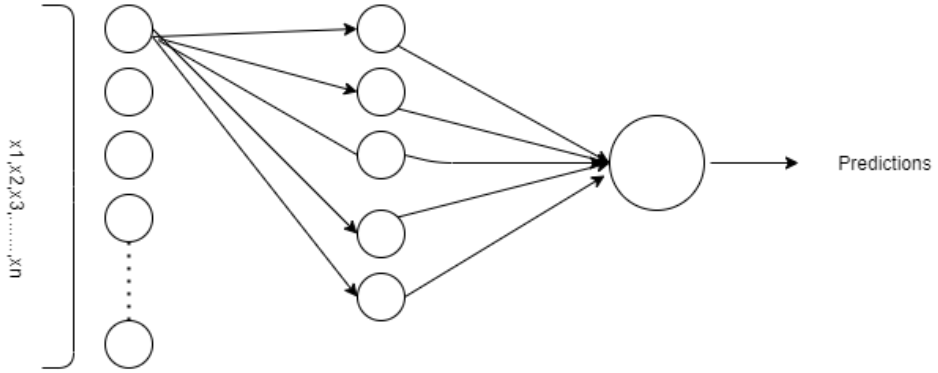
a. Appling existing models on prepared datasets.

1. Random Forest
2. Decision Tree Regression
3. Neural Network Algorithms
4. KNN Model
5. Support Vector Machine (SVM)

b. Preparing Custom model and applying it to prepared datasets

1. Custom model is prepared by combining ANN and Multiple regression.
2. The proposed custom model may have multiple hidden/middle layers.
3. Performance tuning and accuracy of a custom model.
4. Comparative analysis of existing models/algorithms and custom model c.

The model hypothetical architecture is shown below:



The input variables have a complex impact on predictions; hence it requires non-linearity in the model. To introduce non-linearity in the model we need to carefully design the model with multiple layers with non-linear activation functions in between.

Following are the custom model components:

Input layer: The input layer will have multiple input nodes as per the selected variables.

Hidden layers: to introduce non-linearity in the model the hidden layer will be added along with the proper activation function

Activation function: To introduce non-linearity in the model non-linear activation functions are required between the input and subsequent layers. Sigmoid, tanh, and RELU are examples of the non-linear functions

Output layer: Model will have only one node at the output layer which will be used to estimate the recovery value. An activation function at the output layer is required to generate the values greater than 0, RELU can be used to achieve this activation.

Loss function: The loss function is used to compare actuals to predicted values. Here Mean Square Error can be used as a loss function

4.4 Model Training

We will need to train the datasets to run smoothly and see an incremental improvement in the prediction rate. We will check the accuracy of the model by feeding data randomly with the weights for the values that multiply or affect the relationships between the inputs and outputs.

4.5 Evaluation

The accuracy rate of the model will be checked with the real data set. Also, we will find which variable's impact is more to decide the accurate prediction and at all, no impact of variables for prediction represents the data as well as visualize.

4.6 Result

The machine learning model inferring results will be compared in real-life scenarios.

5. Importance of Study/Society Applications:

The machine learning field is continuously evolving. And along with evolution comes a rise in demand and importance. There is one crucial reason why data scientists need machine learning, and that is: 'High-value predictions that can guide better decisions and smart actions in real-time without human intervention.'

Machine learning as technology helps analyze a large volume of data, easing the tasks of data scientists in an automated process, and is gaining a lot of prominence and recognition. Machine learning has changed the way data extraction and interpretation work by involving automatic sets of generic methods that have replaced traditional statistical techniques.

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data over and over, faster and faster is a recent development.

6. Proposed Work Plan:

The proposed study will be completed by the year's plan of work and the target to be achieved. The following table Table1 shows the research plan for accomplishing the prosed study.

Table 1: Proposed Research Plan

Sr. No.	Academic Year	Date		Work Details	Target to be Achieved
		From	To		
1	2021-2022	01-Jan-2022	30-Jun-2022	Reading the core research articles and books Publish one Research Paper	Literature Review will be completed
2		01-Jul-2022	31-Dec-2022	Collect the data Pre-processing the collected data	The data set will be ready
3	2022-2023	01-Jan-2023	30-Jun-2023	Design a Machine Model Publish one Research Paper	Machine Model will be

					ready for testing
4		01-Jul-2023	31-Dec-2023	Test the model on the training data set The accuracy of the result will be tested on real data set.	Machine Model will be implemented and results tested.
5	2023-2024	01-Jan-2024	30-Jun-2024	Thesis writing will be started Publish one Research Paper	The draft thesis will be ready
6		01-Jul-2024	31-Dec-2024	The thesis will be revised and finalized	Work will be submitted.

7. References:

1) Mark A. Kramer, "Nonlinear Principal Component Analysis Using Auto associative neural Networks", AIChE Journal,1991, pp. 233-243, Vol,37, No. 2.

2) Mantas Marclukaitis, Inga Zutautaite, Linas Martisauskas, Benas Joksas, Giedrius, Gecevicius, Athanasios Sfetsos, "Non-Linear regression model for wind turbine power curve", Renewable Energy, 2017, pp. 1-15.

3) Jixiang Sun, Fuping Yang, "Multi-factor Investment Based on TabNet", Journal of Physics: ConferenceSeriess, 2022, pp. 1-5.

4) Jieying Gao, Huan Guo, and Xin Xu, "Multifactor Stock Selection Strategy Based on Machine Learning: Evidence from China", Willey, 2022, pp. 1-17.

5) Fuad Noman, Gamal Alkawsi, Ammar Ahmed Alkahtani, Ali Q. AI- Shetwi, Sieh Kiong Tiong, Nasser Alalwan, Janaka Ekanayake," Multistep Short-term Wind Speed Prediction Using online auto-regressive neural network with groups variable selection", ELSEVIER, 2020, pp. 1221-1229.

6) Roohollah Shirani Faradonbeh, Masoud Monjezi, Danial Jahed Armaghani," Genetic Programming and non-linear multiple regression techniques to predict back break in blasting operation", 2015, springer.

7) Julie E. Shortridge, Stefanie M. Falconi, Benjamin F, Zaitchik and Seth D. Guikema, "Climate, agriculture, and hunger: statistical prediction of undernourishment using nonlinear regression and data –mining technique es, Journal of Applied Statistics, 20 15, pp.1-24.

8) Sambeet Mishra, ChairaBordin, Kota Taharaguchi, Ivo Palu," Comparison of deep learning models for multivariate prediction of time series wind power generation and temperature", ELSEVIER,2019, pp. 273-286.

9) Jan Adamowski, Hiu Fung Chan, Shiv O. Prasher, Bogdan Ozga Zielinski, and Anna Sliusarieva, "Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural net or methods for urban water demand forecasting in Montreal C Canada", Water Resources Research, 2012, Vol.48, pp.1-14.

10) Theresa Hoang Diem Ngo, La Puente, and CA, "The Steps to Follow in a Multiple Regression Analysis", SAS Global Forum 2012, 2012, pp.1-12.

11) Andreas Holzinger, "From Machine Learning to Explainable AI", IEEE DISA 2018 Conference, Kosice, vol.23, 2018, pp.55-66.

12) Konstantinos G. Liakos, "Machine Learning in Agriculture: A Review", Sensors, MDPI, 2018, pp.1-29.

13) Girish L, "Crop yield and Rainfall prediction in Tumkuru district using machine learning", National Conference on Technology for Rural Development 2018, Gubbi, Karnataka 2018, pp.60-65.

14) S Thuankaewsing, "Harvest scheduling algorithm to equalize supplier benefits: A case study from the Thai sugar cane industry", ELSEVIER, 2014, pp.42-55.

15) Manoj G.S., "Prediction and Analysis of Crop Yield using Machine learning Techniques", IJERT ISSN: 2278-0181, 2020, pp.19-23.

16) Y.L.Everingham, C.M.Smith, N .G.Inman-Bamber," Ensemble data mining approaches to forecasting regional sugarcane crop Production", ELSEVIER, 2009, pp.689-696.

17) Phusanisa Charoen-Ung and Mittrapiyanuruk," Sugarcane Yield Grade Prediction Using Random Forest With forwarding Feature Selection and Hyper-Parameter Turning" Springer International Publishing AG, 2019, pp. 33-42.

18) Rajeswari C, Dyuti Basu, and Namita Maurya," Comparative Study of Big data Analytics Tools: R and Tableau", IOP Conference Series: Materials Science and

Engineering, 2017, pp.1-9.

19) Jesper E.VanEngelen, Holger H. Hoos," A Survey On Semi-Supervised Learning",
Springer, 2019, pp. 373-440.

20) Joachim Herbst," A Machine Learning Approach to Workflow Management", Springer, pp.183-194.

21) Sunil Kumar, Vivek Kumar, and K. Sharma, "Sugarcane Yield Forecasting Using Artificial Neural Network Models", International Journal of Artificial Intelligence & Applications, Vol.6, No.5, 2015, pp. 51-68.

22) Jujie Wang, Zhenzhen Zhuang, and Liu Feng," Intelligent Optimization Based Multi-Factor Deep Learning Stock Selection Model and Quantitative Trading Strategy", MDPI mathematics, 2022, pp. 1-19.

23) Devon L. Robertson Wayne Goodridge," Predicting density of serious crime incidents using Multiple –Input Hidden Markov Maximization a posterior model ", Elsevier, 2021, pp. 1-21.

24) Haoxuan Li Xueyan Zhang Ziyan Li Chunyuan Zheng," Overview of machine Learning for Stock Selection Based on Multi-Factor Models", E3S web of Conferences EBLDM, 2020, pp.20221402047.

25) Theresa Hoang Diem Ngo, La Puente, and CA, "The Steps to Follow in a Multiple Regression Analysis", SAS Global Forum 2012, 2012, pp.1-12.

26) Andreas Holzinger, "From Machine Learning to Explainable AI", IEEE DISA 2018 Conference, Kosice, vol.23, 2018, pp.55-66.

Date: 13/04/2022

Place: Latur

(Mr. Surfactant Vishwasrao Patil) (Dr. Renuka R. Londhe)

Research Scholar Research Supervisor / Guide